The Journal of Engineering

The 3rd Asian Conference on Artificial Intelligence Technology (ACAIT 2019)

Deep and CNN fusion method for binaural sound source localisation

eISSN 2051-3305
Received on 14th October 2019
Accepted on 19th November 2019
E-First on 28th July 2020
doi: 10.1049/joe.2019.1207
www.ietdl.org

Shilong Jiang[1], Lulu Wu[2] ✉, Peipei Yuan[2], Yongheng Sun[2], Hong Liu[2]
[1]*PKU-HKUST Shenzhen-Hong Kong Institution, Shenzhen, People's Republic of China*
[2]*Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, People's Republic of China*
✉ *E-mail: susanna@wull.me*

Abstract: In binaural sound source localisation, front–back confusion is often the challenging problem when localising sources in the noisy or reverberant environments. Hence, a novel algorithm fusing deep and convolutional neural network (CNN) is proposed to address this issue. First, joint features, which consist of interaural level differences (ILDs) and cross-correlation function (CCF) within a lag range, are extracted from binaural signals. Second, with the extracted CCF–ILD features, CNN is used for the front–back classification task, while deep neural network is used for azimuth classification task. The front–back features extracted by the CNN can be leveraged as additional information for the sound source localisation task. Also, an angle-loss function is designed to avoid the overfitting problem and to improve the generalisation ability of this method in adverse acoustic conditions. Finally, two branches are concatenated and then followed by an output layer, which generates the posterior probability of azimuth angles, and the azimuth corresponding to the maximum posterior probability is chosen as the direction of sound source. Experimental results demonstrate the effectiveness of the authors' method for front–back decision and azimuth estimation in noisy and reverberant environments.

1 Introduction

Robot auditory system is a natural, convenient, effective and intelligent way for robots to interact with the external world such as sound source localisation, speech enhancement, speech separation, speech recognition etc. [1]. Sound source localisation, as a part of the front-end processing of a robot auditory system, is indispensable for friendly human–robot interaction. As a branch of sound source localisation, binaural sound source localisation (BSSL) can hardly be replaced especially in the fields related to human hearing such as hearing aids, humanoid robots and so on [2–4].

BSSL is to determine the direction of a sound source about a point in space by two microphones mounted on the left and right ears of a dummy head. In 1907, Lord Rayleigh [5] revealed that the incredible ability of a human to localise the sound source is closely related to two principal binaural cues, namely interaural time difference (ITD) and interaural level difference (ILD). ITD is the difference in the time that a sound reaches the left and right ears. ILD is the energy difference between binaural signals caused by the head shadowing effect and diffraction. Therefore, ITD and ILD, which contain source spatial information, are usually extracted from sensor signals for BSSL. The classical method to estimate ITD is the cross-correlation method [6]. Generalised cross-correlation method and its extension are advanced versions of the cross-correlation method, which introduce the cross-power spectrum weighting scheme to improve the robustness in the presence of noise and reverberation. The weights include Roth weight [7], smoothed coherence transform weight [8], phase transform weight [6] etc. ILD is usually estimated by calculating the energy ratio between the left-ear and right-ear signals [9], which is a supplement of ITD especially in high frequency, where the head shadowing effect is obvious.

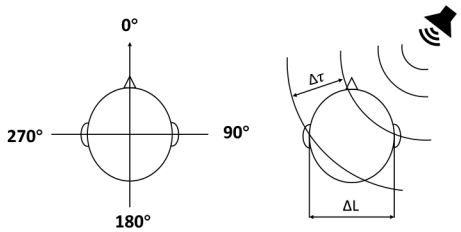

Fig. 1 *BSSL model*

Fig. 1 describes a typical BSSL model in full 360° range. Moreover, the right picture indicates that ITD and ILD are affected by the head according to the scatter theory. However, only depending on ITD [or cross-correlation function (CCF)] or ILD, the robotic auditory systems cannot distinguish the front end from the backplane of sound sources well due to the similarity in the front and rear hemifields. In [10], the author demonstrated that because of the asymmetries between the front and back of the head, ITD and ILD together as one localisation feature could distinguish the front or backside of the sound. However, if there exist reverberation and noise in the acoustic environments, it would introduce the distortion so that ITD would be erroneously estimated by searching the maximum peak of the CCF. Additionally, ILD would also be misestimated due to the presence of noise or reverberation.

With the development of deep learning, some researchers proposed to use CCF directly as the input of deep neural network (DNN) to judge the direction of sound sources because the CCF contains more information than ITD. Ma *et al.* [10] trained DNN for each frequency subband with CCF and ILD and achieved better results than the Gaussian mixture model (GMM) method with ITD and ILD. However, there still exists many front–back confusion.

In this paper, a fused deep and convolutional neural network (DCNN) is proposed for BSSL. First, CCF was calculated from binaural signals and joined with ILD at each frequency subband. Second, the CNN is used to distinguish the front–back hemifields, and DNN is designed to identify the azimuths. Finally, the outputs of these two classifiers are concatenated and followed by a fully connected (FC) layer. To our best knowledge, it is the first time to introduce a front–back classifier as an auxiliary for BSSL. Furthermore, an angle-loss function is proposed to substitute the cross-entropy in training DNN to avoid overfitting. Experiments show that our method performs best even under severe acoustic environment.

2 Fused DCNN system

The received binaural signals emitted from a single source are formulated by convolving the speech source signal with the head-related impulse responses (HRIRs)

$$x_i(n) = h_i(n) \circledast s(n) + v_i(n), \tag{1}$$

where the symbol $\circledast$ denotes the convolution, $i \in \{l, r\}$ denotes the index of the left or right microphone, n denotes the time sampling point within one time frame, $s(n)$ represents a sound source, $h_i(n)$ indicates an HRIR propagating from the source direction to the left or right microphone and $v_i(n)$ denotes the additive noise received by the ith microphone.

Fig. 2 shows a schematic diagram of the fused DCNN system. During training and testing, joint features CCF–ILD are extracted from binaural signals (see more details in Section 2.1). These features are fed into DNN azimuth classifier and CNN front–back classifier. During training, two branches are concatenated and followed by an FC layer. The source direction is determined by the maximum output of DCNN.

2.1 CCF–ILD features

Previous studies have shown that ITD is frequency dependent [11], and so does the ILD because of the head shadowing [12]. Gammatone filter is designed according to the human cochlear sound signal processing, which makes full use of the human ear sound processing characteristic [13]. It is a linear filter consisting of a product of a gamma distribution and a sinusoidal function. Its impulse response is calculated using the equation below:

$$g(t) = At^{m-1}\cos(2\pi ft + \phi)\,\mathrm{e}^{-2\pi bt}. \tag{2}$$

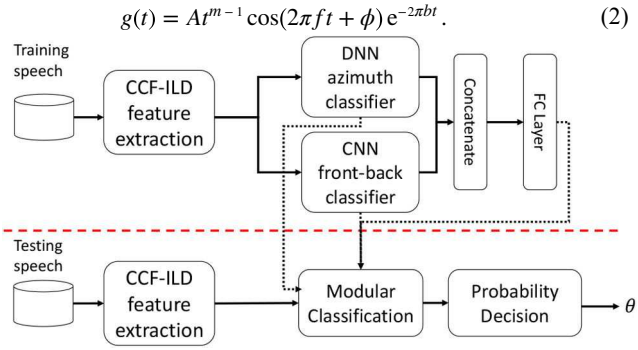

Fig. 2 *Schematic diagram of the fused DCNN system in training and testing phases*

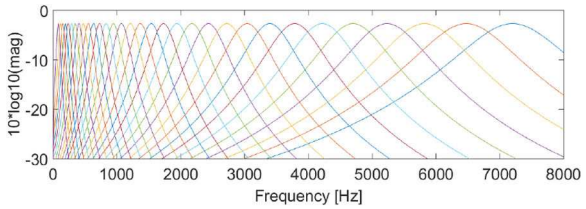

Fig. 3 *Frequency response of gammatone filter*

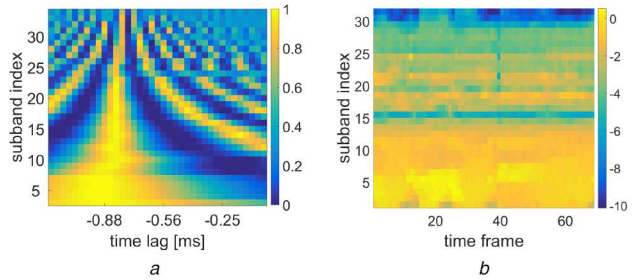

Fig. 4 *Joint CCF–ILD feature extracted from binaural signal, where the source located at azimuth −15° and elevation 0°*
(a) CCF features of 32 filter channels, ***(b)*** ILD features

Therefore, to extract ITD and ILD for different frequencies, a bank of 32 overlapping gammatone filters with the centre frequencies equally distributed for the equivalent rectangular bandwidth scale between 80 Hz and 8 kHz is employed [10].

In (2), m is the order of the filter, b is the bandwidth of the filter, f is the central frequency of the filter, A is the amplitude and t (in seconds) is time. The frequency representation of impulse response is shown in Fig. 3.

The traditional ITD extracted by the CCF method [6] may not be robust in adverse acoustic conditions; therefore, the CCF with lags ranging from −1.1 to 1.1 ms is chosen to replace ITD. The reason we choose this range is that the maximum time differences do not exceed 1 ms according to the distance between two microphones and the speed of sound.

The entire signal is filtered out by a bank of 32 gammatone filters, and the frequency subband index is denoted by k. For each frequency subband, the CCF is calculated as

$$R_{\mathrm{l,\,r}}(k, \tau) = \frac{G_{\mathrm{l,\,r}}(k, \tau)}{\sqrt{G_{\mathrm{l,\,l}}(k, \tau_0) \times G_{\mathrm{r,\,r}}(k, \tau_0)}}, \tag{3}$$

$$G_{i,\,j}(k, \tau) = \sum_n x_i(k, n)x_j(k, n+\tau),\ i, j \in \{l, r\}, \tag{4}$$

where $G_{i,\,j}(k, \tau)$ is the CCF of time delay τ and frequency subband index k between microphones pairs if $i \neq j$; otherwise, it becomes the auto-correlation function of the left or right signal, τ_0 equals to zero. The ILD at each frequency subband is calculated as

$$\mathrm{ILD}(k) = 10\log_{10}\left(\frac{\sum_n x_{\mathrm{r}}(k, n)^2}{\sum_n x_{\mathrm{l}}(k, n)^2}\right). \tag{5}$$

where $x_i(k, n)$ represents the left or right signal at the kth frequency subband. Both CCF and ILD are calculated frame by frame.

For a signal with a sampling rate of 16 kHz, the feature vector within a lag range of ±1.1 ms concludes a 37-dimensional (37D) CCF. Then, supplementing CCF by ILD, a 38D joint feature vector will be extracted from each frequency subband to form a 32×38 feature matrix. It is shown in Fig. 4 the 32 filter channels for CCF features and ILD features, where the sound source located at azimuth −15° and elevation 0°. It can be observed from Fig. 4*a* that the CCF has one local maximum in low frequency, which makes the ITD estimation effective. Besides, there always exist several local maximums in high frequency, which makes it difficult to judge in which local maximum time–frequency fragment the real ITD locates. A different situation can be observed in Fig. 4*b*. The ILD is close to 0 dB in low frequency and thus invalid. That is because the sound wave period is larger than the head diameter, making the sound wave easily around the head. However, the ILD shows strong directional discrimination in high frequencies due to head shadowing effect [14]. Therefore, the combination of CCF and ILD can make the estimation of sound source direction more accurate.

2.2 Fused DCNN

Two neural networks are cascaded in the DCNN model, where DNN is used to determine the direction of the received signal, and CNN is used to assist distinguishing whether the signal is in the front or the back end.

Configuration of DNN: Zheng *et al.* [15] showed that ITD was a function of frequency, and it performed well in the frequency range [500, 2000](Hz). However, the values of ITD and ILD in other frequencies may also slightly affect localisation performance. Therefore, no frequency subbands are excluded in the network inputs. The input layer of DNN contains 1216 nodes, which was obtained by combining the features (CCF and ILD) in all frequency subbands, and the output layer consists of 72 nodes, which represent 72 different directions. DNN consists of three hidden layers with 512 nodes since three hidden layers are enough for

parameter convergence. The rectified linear unit (ReLU) activation function is used in hidden layers.

Configuration of CNN: The CNN model is used to extract more implicit features to identify the front or backside of the source. Local CCF–ILD features show a stronger correlation in adjacent frequency subbands than in all frequency subbands. To strengthen the local relationship between frequency subbands, the CNN model is used to convolve the input features across frequency subbands with a number of convolution kernels of 3×3 size. The CNN model has two convolutional layers with 512 and 1024 feature maps. Each convolutional layer is followed by a ReLU layer and a downsampling pooling layer of size 2×2.

To avoid overfitting, the dropout probability in DCNN is set as 0.2. Both of DNN and CNN are optimised by the Adadelta optimiser and early stopped if there is no lower loss of the validation set within three epochs [16]. More details of the fused two-level DCNN model are shown in Fig. 5. Features of DNN and CNN are concatenated by an FC main output layer of 72 azimuth labels. Joint learning helps to propagate the entire loss backward and update parameters of DNN and CNN so that the mutual information learnt by DNN or CNN can help to improve the other module.

The cross-entropy is usually considered as the loss function in many classification tasks. However, one of its drawbacks is that the classification is too confident even with the noisy input, which usually leads to the overfitting problem. To adapt to unknown environments, a self-entropy loss function is defined for unsupervised training in [17]. It enabled DNN to adapt all the directional signals. As for sound source localisation, the binaural cues are similar in two adjacent directions, so the estimated direction can be accepted within some tolerances. Therefore, we design a smooth angle-loss function by combining cross-entropy and self-entropy

$$J(\Theta) = -(1-\varepsilon)\sum_{o=1}^{N} q_o \log p_o - \varepsilon \sum_{o=1}^{N} p_o \log q_o, \tag{6}$$

where Θ denotes all the network parameters, q_o is the oth probability of the true direction while p_o is the oth output probability of the estimated direction, N is the number of total directions and ε denotes the attention weight of self-entropy and is empirically set to 0.1 in the experiments. If ε equals to 0, the angle-loss function will become the cross-entropy loss function, and if ε equals to 1, it will become the self-entropy loss function. To update all the network parameters Θ, the partial derivative of J to Θ is

$$\frac{\partial J}{\partial \Theta} = \frac{\partial J}{\partial p_o}\frac{\partial p_o}{\partial \Theta}, \quad \frac{\partial J}{\partial p_o} = -(1-\varepsilon)\frac{q_o}{p_o} - \varepsilon \log p_o - \varepsilon. \tag{7}$$

The algorithm is implemented by the toolkit Keras [https://keras.io/]. The angle-loss function is used in DNN's output and DCNN's main output while the cross-entropy function is used in CNN's output. The total loss of DCNN model is the sum of these three losses.

During testing, three probabilities are calculated by the DCNN model for each binaural signal. Assuming that $P_{\text{main}}(\theta)$ denotes the probability of azimuth θ given by the main output, $P_{\text{cnn}}\{\text{front, back}\}$ denotes the probability of front or back end given by the CNN's output and $P_{\text{dnn}}(\theta)$ denotes the probability of azimuth θ given by the DNN's output. Let $\hat{\theta}_{\max}$ denotes the direction corresponding to the maximum $P_{\text{dnn}}(\theta)$ given by

$$\hat{\theta}_{\max} = \arg\max_{\theta} P_{\text{dnn}}(\theta). \tag{8}$$

If $\hat{\theta}_{\max}$ is in the same hemifield as the CNN's output, then $\hat{\theta}_{\max}$ is the final result. Otherwise, if $\hat{\theta}_{\max}$ is in the different hemifield from CNN's output, we consider there is a front–back confusion in estimating the azimuth hat$\theta_{\max}$. So the $\hat{\theta}$ needs to be transformed into the other hemifield by

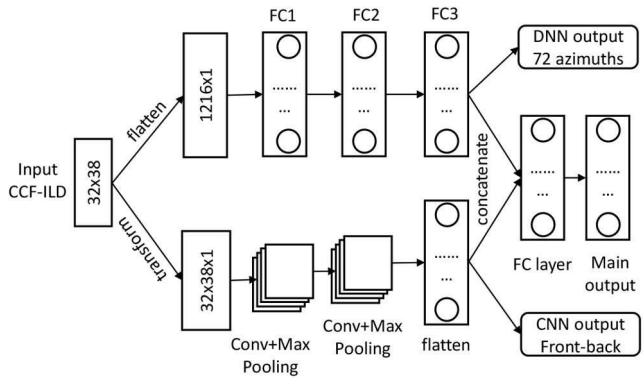

Fig. 5 *Detailed configuration of the fused DCNN*

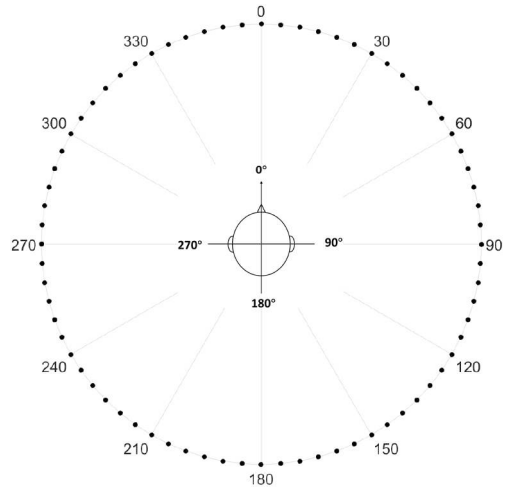

Fig. 6 *Illustration of the binaural setup*

$$\hat{\theta} = \begin{cases} 180 - \theta_{\max}, & \theta_{\max} \in [0, 180] \\ 540 - \theta_{\max}, & \theta_{\max} \in (180, 360). \end{cases} \tag{9}$$

3 Experiments and discussion

3.1 Experimental setup

To evaluate our proposed method, HRIRs measured by the Knowles Electronics Manikin for Acoustic Research [18] are taken to convolve with the source signals. The source signals are selected from the TIMIT database [19]. For training, nine sentences per speaker are uttered by ten men and ten women, i.e. 180 sentences in total. For testing, three sentences per speaker are uttered by three men and three women, i.e. 18 sentences in total. HRIRs of 72 azimuths between 0° and 355° with 5° steps are used in both training and testing. Different dummy heads, which mean different HRIRs, are used in training and testing. A simple illustration of the binaural setup is shown in Fig. 6. The directions in the range of [0°, 90°] and [270°, 355°] are considered in the front end, while the others are in the rear.

To simulate the noisy environment, five kinds of noises {'babble', 'destroyerops', 'factory1', 'white', 'f16'} from NOISEX-92 database [20] are added to the noise-free sensor signals. The first four noises are added to the training set with a signal-to-noise ratio (SNR) in the range of [0:10:30]dB, and the last noise is added to the testing set with SNRs in the range of [−10:10:20]dB. Fig. 7 shows an illustration of these noise signals in the spectrogram sense. The spectrum of 'babble' noise is similar with speech sources, 'destroyerops' noise is a rhythmic wide-band signal, 'factory1' is a kind of irregular noise, 'white' noise is a random signal having equal intensity at different frequencies, the most energy of 'f16' noise is distributed at specific frequencies. Each noise has a different characteristic in the time–frequency domain, which can increase the credibility of our experimental results.

To simulate the room reverberation, five types of binaural room impulse responses (BRIR) are selected from AIR database [21]. Four BRIRs {'booth', 'lecture', 'meeting', 'office'} are convolved with speech signals from the training set. Moreover, {'aula_carolina'} BRIRs are convolved with speech signals from testing set. The average reverberation time $\overline{\mathrm{RT}}_{60}$ for each room is shown in Table 1.

The accuracy of front–back confusion is measured by the percentage of the number of correct classification to the total number of binaural signals. The accuracy of direction of arrival (DOA) estimation is also evaluated by the percentage of the number of correctly estimated azimuths to the total number of binaural signals in terms of the tolerances of 0°, 5° and 10°, which is defined as

$$\mathrm{Acc}(\%) = \frac{N_{|\hat{\theta}-\theta| \le T}}{N_{\mathrm{total}}} \times 100\% . \tag{10}$$

where $\hat{\theta}$ denotes the estimated DOA, θ denotes the true DOA and T corresponds to the aforementioned three tolerances.

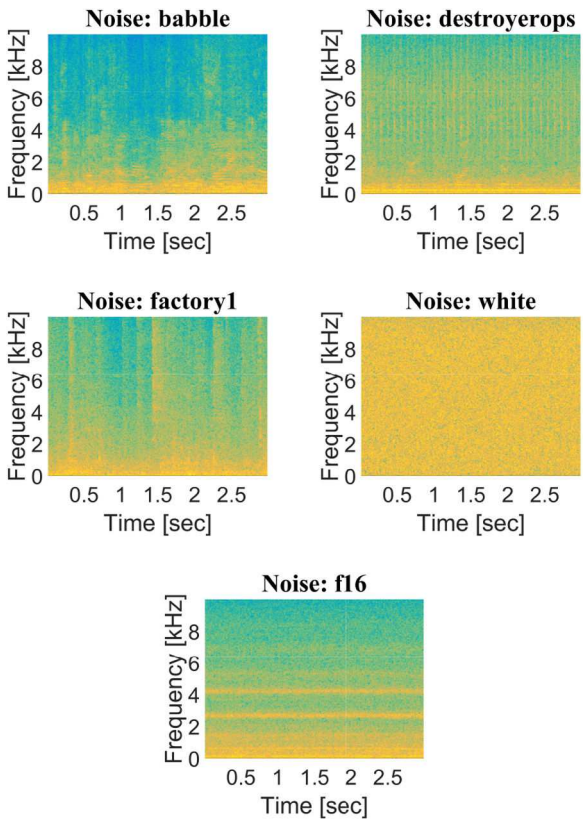

Fig. 7 *Spectrum of five noise signals used in experiments*

3.2 Localisation performance

The first experiment presents the localisation accuracy of different methods in a noiseless, noisy and reverberant environment. Each method is evaluated within tolerances of 0°, 5° and 10°, respectively. In Table 2, avg denotes the average accuracy. Four baseline models: DNN (Freq.Indep.) [10], DNN (cross-entropy), DNN (angle loss) and CNN (angle loss) are compared with our fused DCNN model. DNN (Freq.Indep.) is trained for 32 frequency subbands independently. To compare cross-entropy with angle-loss function, DNN (cross-entropy) has the same configuration with DNN (angle loss), except the loss function. The configuration of CNN (angle loss) is the same as the one of our CNN front–back classifier, except for the output. The outputs of CNN (angle loss) are the 72 probabilities of azimuths. All the above models are trained in noisy and reverberant environments, and then tested in noiseless, noisy and reverberant environments.

Table 2 shows that DCNN model performs the best over three scenarios in terms of the average accuracy. It improves the accuracy of more than 5% over the second maximum. In the noiseless environment, the localisation accuracy of four models is >95%, expect DNN (Freq.Indep.) within tolerance 0°. Additionally, the localisation accuracy of DNN models with different loss functions, DNN (cross-entropy) and DNN (angle loss), achieves 100% within tolerances of three kinds of degrees in the noiseless environment, while the performance of CNN (angle loss) is not as good as the one of DNN. This phenomenon indicates that the DNN model in our proposed method is more suitable than CNN to locate azimuths. Moreover, the same phenomenon can be observed in the noisy environment. In the noisy environment under SNR = −10 dB, DCNN model shows the best accuracy within tolerances of 10° while DNN (Freq.Indep.) shows the worst results. As for the comparison between cross-entropy and angle-loss function, DNN (angle loss) presents better results in the noisy environment, but worst results in reverberation than DNN (cross-entropy). That is because the received signals may come from different directions due to the room reflection, the true direction may occur with second maximum probability. Moreover, it is distinctly confirmed that fused DCNN model improves the localisation accuracy of more than 11% over the DNN (Freq.Indep.) model in the reverberation within tolerances 0°. The fused DCNN model can take advantage of both DNN and CNN so that it generalises well in noisy and reverberant environments.

To evaluate the localisation performance in the noisy environments under different SNRs, the localisation accuracy of the aforementioned methods within tolerances of 10° is depicted in Fig. 8, SNRs are in the range of [−10:10:20] dB. The results demonstrate that DCNN model is robust in noisy environments. However, the binaural cues are dramatically deteriorated by the noise under SNR lower than −10 dB.

3.3 Front–back classifier

The second experiment is to evaluate the performance of front–back classifiers. Fig. 9 shows the front–back classification

Table 1 Reverberation time of different rooms

Room	Booth	Lecture	Meeting	Office	Aula Carolina
$\overline{\mathrm{RT}}_{60}$	0.12 s	0.78 s	0.23 s	0.43 s	5.16 s

Table 2 Localisation accuracy in different environments

	Noiseless			Noisy (SNR = −10 dB)			Reverberant			Average
Tolerance	=0°, %	≤5°, %	≤10°, %	=0°, %	≤5°, %	≤10°, %	=0°, %	≤5°, %	≤10°, %	Average, %
DNN (Freq.Indep.) [10]	81.13	98.84	99.83	67.13	69.91	69.91	43.43	61.11	77.78	74.34
DNN (cross-entropy)	100	100	100	93.52	93.52	93.52	41.41	41.41	41.41	<u>78.31</u>
DNN (angle loss)	100	100	100	95.83	95.83	95.83	23.74	23.74	23.74	73.19
CNN (angle loss)	97.57	100	100	92.59	92.59	92.59	36.87	40.40	40.91	77.06
DCNN	99.65	100	100	94.91	95.37	96.30	54.55	55.05	55.05	**83.43**

Bold values indicates underline the average result of only using DNN with cross-entropy loss, and wo want to highlight our result compared with DNN (Freq.Indep.) [10] and DNN (angle loss). The bold indicates the result using the proposed method and we want to highlight our result compared with all the methods.

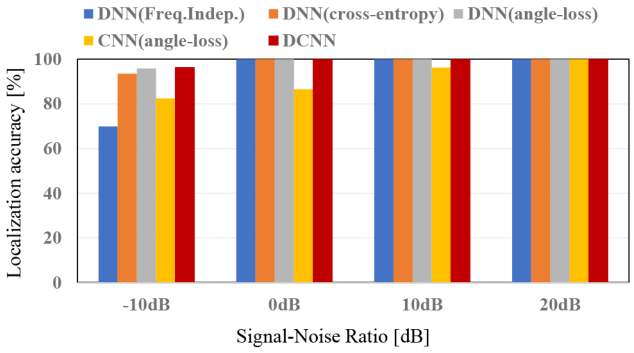

Fig. 8 *Localisation accuracy in noisy environment with different SNRs*

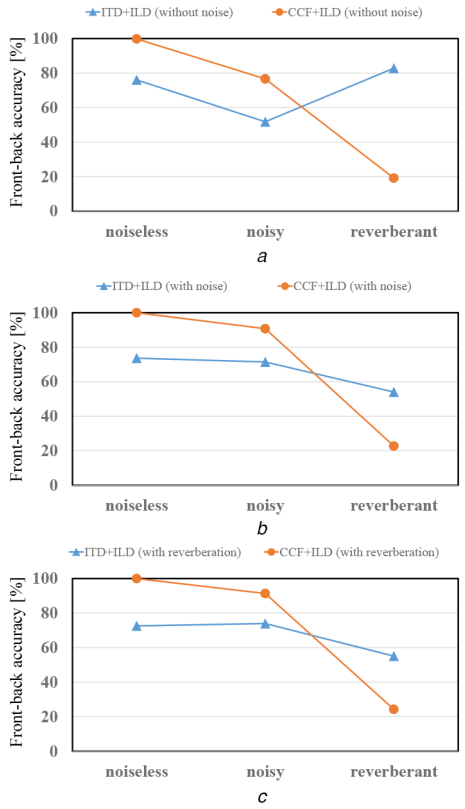

Fig. 9 *Front–back accuracy under different training and testing acoustic conditions. The DNN models using CCF–ILD or ITD–ILD features are trained in*

(a) Noiseless, ***(b)*** Noisy, ***(c)*** Reverberant environment and tested in three environments: noiseless, noisy and reverberant

accuracy of DNN models with ITD–ILD or CCF–ILD features. To testify the robustness of ITD–ILD and CCF–ILD features, the DNN model consisting of 2 hidden layers with 128 nodes takes ITD–ILD or CCF–ILD features as inputs, respectively. It can be observed from Fig. 9 that CCF–ILD features are more robust than ITD–ILD features in noiseless and noisy environments, but worst in reverberation. It is because that the overall frequency distribution is influenced by the reverberation, but ITD–ILD integrates all frequencies non-linearly without filtering, which enables to capture more accurate information in reverberation. To evaluate the classification models, CNN is used to convolve with CCF–ILD non-linearly. Additionally, in all scenarios, DNN models trained with noise-free CCF–ILD features perform the worst. The main reason is that the noises we added to binaural signals have damaged binaural information, which was captured by DNN. This experiment proves that CCF–ILD features outperform ITD–ILD features in most environments, so we will use CCF–ILD features to train the front–back classifier in the following experiment.

To evaluate the classification ability of different networks such as DNN, our front–back classifier CNN is compared with DNN. Fig. 10 shows the front–back classification accuracies of CNN and

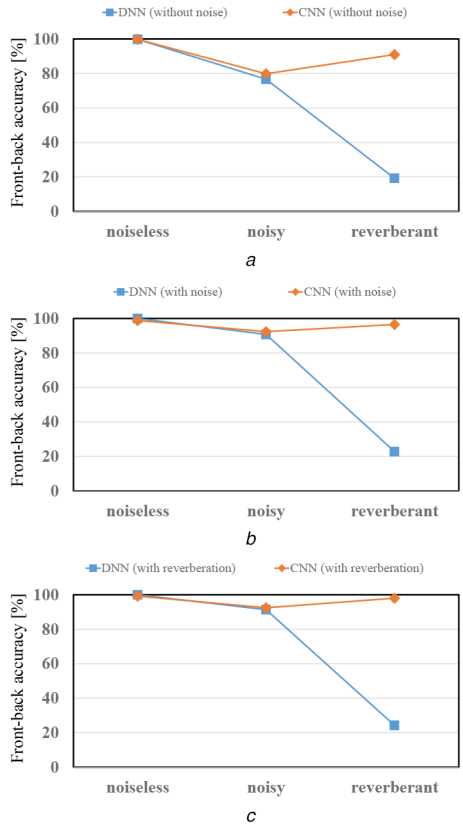

Fig. 10 *Front–back accuracy under different training and testing acoustic conditions. The DNN model and CNN model are trained with CCF–ILD features in*

(a) Noiseless, ***(b)*** Noisy, ***(c)*** Reverberant environment and tested in noiseless, noisy and reverberant environments

DNN models trained with CCF–ILD features. These models are trained in noiseless, noisy and noisy reverberant environments, respectively. It can be seen from Fig. 10 that the CNN model keeps the highest front–back classification accuracy in all scenarios. The front–back accuracy of CNN model trained with CCF–ILD features is above 80% in the three tested environments.

Furthermore, it can be observed that CNN model outperforms the DNN model using same features by more than 70% accuracy in the reverberant environment, while the DNN model and CNN model perform comparably in the noiseless and noisy environment. This experiment distinctly confirms that CNN can extract more discriminative binaural features than DNN in the front–back classification, especially in the strongly reverberant environments. In the following experiment, we will use the CNN model as the front–back classifier and fuse it with DNN azimuth classifier.

To evaluate the effect of front–back confusion, Table 3 describes the accuracy of front–back confusion in different environments. It can be observed that CNN (angle loss) has the highest accuracy when compared with DNN (cross-entropy) and DNN (angle loss) models overall environments within tolerances of 0°, 5° and 10°, which indicates that CNN model in the proposed method is more suitable than DNN to distinguish the front from back. All of these models show almost 100% front–back confusion accuracy in the noiseless environment. In the noisy and reverberant environments, the DCNN model reduces front–back confusion significantly. This is attributed to the strong front–back classification ability of the CNN.

4 Conclusions

This paper presents a novel algorithm fusing DCNN for BSSL. The front–back classifier CNN can generate robust front–back features by convolving kernels on the CCF–ILD features, serving as the additional procedure for sound source localisation task and reducing the front–back error. By jointly exploiting DNN and CNN to construct the fused DCNN model, this system can alleviate the

Table 3 Front–back classification accuracy in different environments

	Noiseless			Noisy (SNR = −10 dB)			Reverberant		
Tolerance	=0°, %	≤5°, %	≤10°, %	=0°, %	≤5°, %	≤10°, %	=0°, %	≤5°, %	≤10°, %
DNN (Freq.Indep.) [10]	100	99.83	99.83	100	90.28	86.11	97.47	82.83	82.32
DNN (cross-entropy)	100	100	100	100	95.37	94.44	78.28	71.21	71.21
DNN (angle loss)	100	100	100	100	96.76	96.76	81.31	73.74	64.65
CNN (angle loss)	100	100	100	100	97.22	97.22	84.85	81.82	81.82
DCNN	100	100	100	99.07	99.07	99.07	96.46	93.43	93.43

localisation error caused by font-back confusion. In addition, to avoid the overfitting problem during the training phase, the angle-loss function is employed instead of cross-entropy, and it shows better performance in noisy environment. All the aforementioned experimental results show that by exploiting fused DCNN (in this way), the generalised robustness can be improved under conditions, where the noise and reverberation are present.

However, this paper only focuses on the binaural localisation of a single sound source, and we will introduce multiple sound sources under complex acoustic conditions in future work.

5 Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC nos. 61673030 and U1613209), the Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No. ZDSYS201703031405467).

6 References

[1] Wang, D., Brown, G.J.: '*Computational auditory scene analysis: principles, algorithms, and applications*' (Wiley-IEEE Press, 2006)

[2] Lyon, L.: 'A computational model of binaural localization and separation'. IEEE Int. Conf. Acoustics Speech and Signal Processing, 1983, pp. 1148–1151

[3] Jeub, M., Schafer, M., Esch, T., *et al.*: 'Model-based dereverberation preserving binaural cues', *IEEE Trans. Audio Speech, Lang. Process.*, 2010, **18**, (7), pp. 1732–1745

[4] Nakadai, K., Matsuura, D., Okuno, H.G., *et al.*: 'Applying scattering theory to robot audition system: robust sound source localization and extraction'. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, Las Vegas, NV, USA, 2003, pp. 1147–1152

[5] Lord Rayleigh, O.M.P.R.S.: 'XII. on our perception of sound direction', *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, 1907, **13**, (74), pp. 214–232

[6] Knapp, C., Carter, G.: 'The generalized correlation method for estimation of time delay', *IEEE/ACM Trans. Audio Speech, Lang. Process.*, 1976, **24**, (4), pp. 320–327

[7] Roth, P.R.: 'Effective measurements using digital signal analysis', *IEEE Spectr.*, 1971, **8**, (4), pp. 62–70

[8] Carter, G.C., Nuttall, A.H., Cable, P.G.: 'The smoothed coherence transform', *Proc. IEEE*, 1973, **61**, (10), pp. 1497–1498

[9] May, T., Van de Par, S., Kohlrausch, A.: 'A probabilistic model for robust localization based on a binaural auditory front end', *IEEE/ACM Trans. Audio Speech, Lang. Process.*, 2011, **19**, (1), pp. 1–13

[10] Ma, N., May, T., Brown, G.J.: 'Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments', *IEEE/ACM Trans. Audio Speech, Lang. Process.*, 2017, **25**, (12), pp. 2444–2453

[11] Kuhn, G.F.: 'Model for the interaural time differences in the azimuthal plane', *J. Acoust. Soc. Am.*, 1977, **62**, (1), pp. 157–167

[12] Blauert, J.: '*Spatial hearing: the psychophysics of human sound localization*' (MIT Press, Cambridge, MA, USA, 1997)

[13] Patterson, R.D., Holdsworth, J., Allerhand, M.: 'Auditory models as preprocessors for speech recognition', '*The auditory processing of speech: from auditory periphery to words*' (1992), pp. 67–89

[14] Argentieri, S., Dans, P., Soures, P.: 'A survey on sound source localization in robotics: from binaural to array processing methods', *Comput. Speech Lang.*, 2015, **34**, (1), pp. 87–112

[15] Zheng, C., Schwarz, A., Kellermann, W., *et al.*: 'Binaural coherent-to-diffuse ratio estimation for dereverberation using an ITD model'. European Signal Processing Conf., Nice, France, 2015, pp. 1048–1052

[16] Zeiler, M.D.: 'ADADELTA: an adaptive learning rate method', arXiv preprint arXiv:1212.5701, 2012

[17] Takeda, R., Komatani, K.: 'Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization'. IEEE Int. Conf. Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 2017, pp. 2217–2221

[18] Wierstorf, H., Geier, M., Spors, S.: 'A free database of head-related impulse response measurements in the horizontal plane with multiple distances'. Audio Engineering Society Convention 130, 2011

[19] Garofolo, J.S., Lamel, L.F., Fisher, W.M., *et al.*: 'DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1–1.1'. NASA STI/Recon Technical Report N, 1993, 93

[20] Varga, A., Steeneken, H.J.: 'Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems', *Speech Commun.*, 1993, **12**, (3), pp. 247–251

[21] Jeub, M., Schafer, M., Vary, P.: 'A binaural room impulse response database for the evaluation of dereverberation algorithms'. Int. Conf. Digital Signal Processing, 2009, pp. 1–5